DOCUMENT RESUME

ED 306 308 TM 013 253

AUTHOR Federico, Pat-Anthony
TITLE Computer-Based and Paper-Based Measurement of Recognition Performance.
INSTITUTION Navy Personnel Research and Development Center, San Diego, Calif.
REPORT NO NPRDC-TR-89-7
PUB DATE Mar 89
NOTE 31p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Armed Forces; Comparative Analysis; *Computer Assisted Testing; Discriminant Analysis; *Flight Training; *Military Personnel; Multivariate Analysis; Postsecondary Education; Professional Training; *Recognition (Psychology); Test Format; *Test Reliability; Test Validity
IDENTIFIERS *Paper and Pencil Tests; Radar Intercept Observers

ABSTRACT

To determine the relative reliabilities and validities of paper-based and computer-based measurement procedures, 83 male student pilots and radar intercept officers were administered computer and paper-based tests of aircraft recognition. The subject matter consisted of line drawings of front, side, and top silhouettes of aircraft. Reliabilities for both modes of testing were estimated by deriving internal consistency indices, using an odd/even item split. Prior to testing, subjects learned to recognize the aircraft silhouettes using two media: (1) paper-based form structured as a study guide; and (2) computer-based form using FLASH IVAN in the training mode. A stepwise multiple discriminant analysis was performed to determine how well the two testing modes distinguished among two groups of subjects expected to differ in their recognition of aircraft silhouettes. Computer-based and paper-based measures were not significantly different in reliability or internal consistency. The paper-based measure of average degree of confidence in recognition judgments was more reliable than the computer-based measure. The average degree of confidence measured by the two modes was more equivalent than the measures of recognition test scores. The discriminative validities of the two measures were about the same for distinguishing groups above or below the mean average curriculum grade. Using the pooled within-groups correlations between the discriminant function and computer-based or paper-based measures, the former had superior discriminative validity than the latter. Statistics associated with the canonical correlation suggested that the predictive validity of computer-based measures approximates that of paper-based measures. Four tables present study data. A 54-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Navy Personnel Research and Development Center

San Diego, CA 92152-6800 TR 89-07 March 1989

ED306308

Computer-based and Paper-based Measurement of Recognition Performance

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

PAT-ANTHONY FEDERICO

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Approved for public release; distribution is unlimited.

TM013253

BEST COPY AVAILABLE

NPRDC TR 89-7 March 1989

COMPUTER-BASED AND PAPER-BASED MEASUREMENT OF RECOGNITION PERFORMANCE

Pat-Anthony Federico
Navy Personnel Research and Development Center

Reviewed and approved by
E. G. Aiken

Released by
B. E. Bacon
Captain, U. S. Navy
Commanding Officer
and
J. S. McMichael
Technical Director

Approved for public release;
distribution is unlimited.

Navy Personnel Research and Development Center
San Diego, California 92152-6800

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NPRDC TR 89-7	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Navy Personnel Research and Development Center 6b. OFFICE SYMBOL (If applicable) Code 15	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) San Diego CA 92152-6800	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Office of Naval Technology 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) Washington DC 20350	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 62233N PROJECT NO: RF62-522 TASK NO: 801-013 WORK UNIT ACCESSION NO: 03.04
11. TITLE (Include Security Classification) COMUTER-BASED AND PAPER-BASED MEASUREMENT OF RECOGNITION PERFORMANCE	
12. PERSONAL AUTHOR(S) Federico, Pat-Anthony	
13a. TYPE OF REPORT Technical Report 13b. TIME COVERED FROM 86 Oct TO 87 Dec 14. DATE OF REPORT (Year, Month, Day) 1989 March 15. PAGE COUNT 27	
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Computer-based testing, measurement, assessment; modes, manners of testing; test-item administration.
05	09		

19. ABSTRACT

Eighty-three subjects were administered computer-based and paper-based tests to assess recognition of aircraft silhouettes in order to determine the relative reliabilities and validities of these two measurement modes. Estimates of internal consistencies, equivalences, and discriminative and predictive validities were computed. It was established that (a) computer-based and paper-based measures of recognition test score were not significantly different in reliability or internal consistency, (b) the paper-based measure of average degree of confidence in recognition judgments was more reliable or internally consistent than the computer-based measure, (c) computer-based and paper-based measures of average degree of confidence were more equivalent than these measures of recognition test score, (d) according to two sets of criteria, the discriminant coefficients and F-ratios and corresponding means, the discriminative validities of computer-based and paper-based measures were about the same for distinguishing groups above or below mean average curriculum grade, (e) according to another set of criteria, the pooled within-groups correlations between the discriminant function and computer-based and paper-based measures, the former had superior discriminative validity than the latter, and (f) statistics associated with the canonical correlation suggested the predictive validity of computer-based measures approximates that of paper-based measures.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Pat-Anthony Federico	22b. TELEPHONE (Include Area Code) (619) 553-7777 22c. OFFICE SYMBOL Code 15

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

FOREWORD

This research was performed under exploratory development work unit RF63-522-801-013-03.04, Testing Strategies for Operational Computer-based Training, under the sponsorship of the Office of Naval Technology, and advanced development project Z1772-ET006, Computer-Based Performance Testing, under the sponsorship of Deputy Chief of Naval Operations (Manpower, Personnel, and Training). The general goal of this development is to create and evaluate computer-based simulations of operationally oriented tasks to determine if they result in better assessment of student performance than more customary measurement methods.

The results of this study are primarily intended for the Department of Defense training and testing research and development community.

B. E. BACON
Captain, U.S. Navy
Commanding Officer

J. S. MCMICHAEL
Technical Director

SUMMARY

Background

The literature regarding computer-based assessment is contradictory and inconclusive: Many benefits may be obtained from computerized testing. Some of these may be related to attitudes and assumptions associated with the use of novel media or innovative technology per se. However, and just as readily, potential problems may result from the employment of computer-based measurement. Differences between this mode of assessment and traditional testing techniques may, or may not, impact upon the reliability and validity of measurement. Notably absent from this literature are studies that have compared these testing characteristics of computer-based assessment with customary measurement methods for estimating recognition performance.

Problem

Many student assessment schemes which are currently used in Navy training are suspected of being insufficiently accurate or consistent. If true, this could result in either overtraining, which increases costs needlessly, or undertraining, which culminates in unqualified graduates being sent to the fleets.

Objective

The specific objective of this research was to compare the reliability and validity of a computer-based and a paper-based procedure for assessing recognition performance.

Method

A computer-based and paper-based test were developed to assess recognition of Soviet and non-Soviet aircraft silhouettes. These tests were administered to 83 student pilots and radar intercept officers from the F-14 Fleet Replacement Squadron, VF-124, NAS Miramar. All volunteered to participate in this study. After the subjects received the paper-based test, they were immediately given the computer-based test. It was assumed that a subject's state of recognition knowledge was the same during the administration of both tests.

Reliabilities for both modes of testing were estimated by deriving internal consistency indices using an odd-even item split. These estimates were adjusted by employing the Spearman-Brown Prophecy Formula. Reliability estimates were calculated for test score, average degree of confidence, and average response latency for the computer-based test; reliability estimates were calculated for test score and average degree of confidence only for the paper-based test. None was computed for average response latency since this was not measured for the paper-based test. Equivalences between these two modes of assessment were estimated by Pearson-product-moment correlations for total test score and average degree of confidence.

In order to derive discriminant validity estimates, subjects were placed into two groups according to whether or not their performance through the squadron's curriculum was above or below the mean average grade for this sample. A stepwise multiple discriminant analysis, using Wilks' criterion for including and rejecting variables, and their associated statistics were computed to ascertain how well computer-based and paper-based measures distinguished among the defined groups expected to differ in their recognition of aircraft silhouettes. Predictive validity indices were obtained by computing a canonical analysis between computer-based and paper-based recognition measures and subjects' test scores for each phase of the curriculum.

Results

It was demonstrated that (a) computer-based and paper-based measures of recognition test score were not significantly different in reliability or internal consistency, (b) the paper-based measure of average degree of confidence in recognition judgments was more reliable or internally consistent than the computer-based measure, (c) computer-based and paper-based measures of average degree of confidence were more equivalent than these measures of recognition test score, (d) according to two sets of criteria, the discriminant coefficients and F-ratios and corresponding means, the discriminative validities of computer-based and paper-based measures were about the same for distinguishing groups above or below mean average curriculum grade, (e) according to another set of criteria, the pooled within-groups correlations between the discriminant function and computer-based and paper-based measures, the former had superior discriminative validity than the latter, and (f) statistics associated with the canonical correlation suggested the predictive validity of computer-based measures approximates that of paper-based measures.

Discussion

This study established that the relative reliability of computer-based and paper-based measures depends upon the specific criterion assessed. That is, regarding the recognition test score itself, it was found that computer-based and paper-based measures were not significantly different in reliability or internal consistency. However, regarding the average degree of confidence in recognition judgments, it was found that the paper-based measure was more reliable or internally consistent than its computer-based counterpart. The extent of the equivalence between these two modes of measurement was contingent upon particular performance criteria. It was demonstrated that the equivalence of computer-based and paper-based measures of average degree of confidence was greater than that for recognition test score. The relative discriminative validity of computer-based and paper-based measures was dependent upon the specific statistical criteria selected. The discriminant coefficients, F-ratios, and corresponding means indicated that the validities of computer-based and paper-based measures were about the same for distinguishing groups above or below mean average curriculum grade. However, according to another set of criteria, the pooled within-groups correlations between the discriminant function and computer-based and paper-based measures, the former had superior validity than the latter. Also, according to the statistics

associated with the canonical correlation, this study demonstrated that the predictive validity of computer-based measures approximates that of paper-based measures. The results of this research supported the findings of some studies, but not others. As was discussed, the reported literature on this subject is contradictory and inconclusive.

RECOG, the computer-based system for assessing recognition performance, together with the Soviet and non-Soviet aircraft silhouette database, is referred to as FLASH IVAN. This system is currently being used to augment the teaching and testing of this subject matter in VF-124. RECOG was designed and developed with generalizability (i.e., independence of subject-matter domain) and transferability (i.e., capable of readily running on different computer systems) in mind as was the Computer-Based Educational Software System (CBESS) (Brandt, 1987; Brandt, Gay, Othmer & Halff, 1987). CBESS consists of a number of component quizzes such as JEOPARDY, TWENTY QUESTIONS, and CONSTRAINT. Since the time that RECOG and FLASH IVAN were developed and evaluated, two other tests, FLASH and PICTURE, were added to CBESS. In terms of their function, these additional quizzes are similar to RECOG and FLASH IVAN. However, these more recently produced tests are written in the "C" programming language for the Navy standard microcomputer, Zenith Z-248. Consequently, since TERAK computers are no longer being produced, the company went out of business, FLASH and PICTURE can be perceived as replacements for RECOG and FLASH IVAN.

Recommendations

Based upon the findings of this study, the following actions are recommended:

(a) Commander, Naval Air Force, U.S. Pacific Fleet, use FLASH and PICTURE to supplement the training and testing of fighter and other crew members to recognize Soviet and Non-Soviet silhouettes.

(b) Chief of Naval Operations, Total Force Training and Education, fund the evaluation and seek implementation of FLASH and PICTURE in other content areas or subject-matter domains (e.g., ship silhouettes, electronic schemata, human anatomy) to ascertain the universality of the validity and reliability results established in this reported research.

CONTENTS

	Page
INTRODUCTION	1
Background	1
Problem	2
Objective	3
METHOD	3
Subjects	3
Subject Matter	3
Computer-Based Assessment	4
Paper-Based Assessment	5
Procedure	5
RESULTS	6
Reliability and Equivalence Estimates	6
Discriminative Validity	7
Predictive Validity	8
DISCUSSION	8
RECOMMENDATIONS	11
REFERENCES	13
APPENDIX--TABLES OF RELIABILITY AND VALIDITY ESTIMATES	A-0
DISTRIBUTION LIST	

INTRODUCTION

Background

The consequences of computer-based assessment on examinees' performance are not obvious. The investigations that have been conducted on this topic have produced mixed results. Some studies (Serwer & Stolurow, 1970; Johnson & Mihal, 1973) demonstrated that testees do better on verbal items given by computer than paper-based; however, just the opposite was found by other studies (Johnson & Mihal, 1973; Wildgrube, 1982). One investigation (Sachar & Fletcher, 1978) yielded no significant differences resulting from computer-based and paper-based modes of administration on verbal items. Two studies (English, Reckase & Patience, 1977; Hoffman & Lundberg, 1976) demonstrated that these two testing modes did not effect performance on memory retrieval items. Sometimes (Johnson & Mihal, 1973) testees do better on quantitative tests when computer given, sometimes (Lee, Moreno, & Sympson, 1984) they do worse, and other times (Wildgrube, 1982) it may make no difference. Other studies have supported the equivalence of computer-based and paper-and-paper administration (Elwood & Griffin, 1972; Hedl, O'Neil, & Hansen, 1973; Kantor, 1988; Lukin, Dowd, Plake, & Kraft, 1985). Some researchers (Evan & Miller, 1969; Koson, Kitchen, Kochen, & Stodolosky, 1970; Lucas, Mullin, Luna, & McInroy, 1977; Lukin, Dowd, Plake, & Kraft, 1985; Skinner & Allen, 1983) have reported comparable or superior psychometric capabilities of computer-based assessment relative to paper-based assessment in clinical settings.

Investigations of computer-based administration of personality items have yielded reliability and validity indices comparable to typical paper-based administration (Katz & Dalby, 1981; Lushene, O'Neil, & Dunn, 1974). No significant differences were found in the scores of measures of anxiety, depression, and psychological reactance due to computer-based and paper-based administration (Lukin, Dowd, Plake, & Kraft, 1985). Studies of cognitive tests have provided inconsistent findings with some (Hitti, Riffer, & Stuckles, 1971; Rock & Nolen, 1982) demonstrating that the computerized version is a viable alternative to the paper-based version. Other research (Hansen & O'Neil, 1970; Hedl, O'Neil, & Hansen, 1973; Johnson & White, 1980; Johnson & Johnson, 1981), though, indicated that interacting with a computer-based system to take an intelligence test could elicit a considerable amount of anxiety which could affect performance.

Regarding computerized adaptive testing (CAT), some empirical comparisons (McBride, 1980; Sympson, Weiss, & Ree, 1982) yielded essentially no change in validity due mode of administration. However, test-item difficulty may not be indifferent to manner of presentation for CAT (Green, Bock, Humphreys, Linn, & Reckase, 1984). When going from paper-based to computer-based administration, this mode effect is thought to have three aspects: (a) an overall mean shift where all items may be easier or harder, (b) an item-mode interaction where a few items may be altered and others not, and (c) the nature of the task itself may be changed by computer administration. These inconsistent results of mode, manner, or medium of testing may be due to differences in methodology, test content, population tested, or the design of the study (Lee, Moreno & Sympson, 1984).

With computer costs coming down and peoples' knowledge of these systems going up, it becomes more likely economically and technologically that many benefits can be gained from their use. A direct advantage of computer-based testing is that individuals can respond to items at their own pace, thus producing ideal power tests. Some indirect advantages of computer-based assessment are increased test security, less ambiguity about students' responses, minimal or no paperwork, immediate scoring, and automatic records keeping for item analysis (Green, 1983a, 1983b). Some of the strongest support for computer-based assessment is based upon the awareness of faster and more economical measurement (Elwood & Griffin, 1972; Johnson & White, 1980; Space, 1981). Cory (1977) reported some advantages of computerized over paper-based testing for predicting on job performance.

Ward (1984) stated that computers can be employed to augment what is possible with paper-based measurement (e.g, to obtain more precise information regarding a student than is likely with more customary measurement methods) and to assess additional aspects of performance. He enumerated and discussed potential benefits that may be derived from employing computer-based systems to administer traditional tests. Some of these are as follows: (a) individualizing assessment, (b) increasing the flexibility and efficiency for managing test information, (c) enhancing the economic value and manipulation of measurement databases, and (d) improving diagnostic testing. Millman (1984) claimed to agree with Ward, especially regarding the ideas that computer-based measurement encourages individualizing assessment, designing software within the context of cognitive science, and limiting computer-based assessment is not so much hardware inadequacy but incomplete comprehension of the processes intrinsic to testing and knowing per se (Federico, 1980).

As is evident, the literature regarding computer-based assessment is contradictory and inconclusive: Many benefits may be obtained from computerized testing. Some of these may be related to attitudes and assumptions associated with the use of novel media or innovative technology per se. However, and just as readily, potential problems may result from the employment of computer-based measurement. Differences between this mode of assessment and traditional testing techniques may, or may not, impact upon the reliability and validity of measurement. Notably absent from this literature are studies that have compared these testing characteristics of computer-based assessment with customary measurement methods for assessing recognition performance.

Problem

Many student assessment procedures which are currently used in Navy training are suspected of being insufficiently accurate or consistent. If true, this could result in overtraining, which increases costs needlessly, or undertraining, which culminates in unqualified graduates being sent to the fleet commands. Many of the customary methods for measuring performance either on the job or in the classroom involve instruments which are primarily paper-based in nature (e.g., check lists, rating scales, critical incidences; and multiple-choice, completion, true-false, and matching formats). A number of deficiencies exist with these traditional testing techniques such as (a) biased items are generated by different individuals, (b) item writing procedures are

usually obscure, (c) there is a lack of objective standards for producing tests, (d) item content is not typically sampled in a systematic manner, and (e) there is usually a poor relationship between what is taught and test content.

What is required is a theoretically and empirically grounded technology of producing procedures for testing which will correct these faults. One promising approach employs computer technology. However, very few data are presently available regarding the psychometric properties of testing strategies using this technology. Data are needed concerning the accuracy, consistency, sensitivity, and fidelity of these computer-based assessment schemes compared to more traditional testing techniques.

Objective

The specific objective of this research was to compare the reliability and validity of a computer-based and a paper-based procedure for assessing recognition performance.

METHOD

Subjects

The subjects were 83 male student pilots and radar intercept officers (RIOs) from the Fleet Replacement Squadron, VF-124, NAS Miramar, who volunteered to participate in this study. This squadron trains crew members to fly the F-14 fighter as well as make intercepts using its many complex systems. One of the major missions of the F-14 is to protect carrier-based naval task forces against antiship, missile-launching, threat bombers. This part of the F-14's mission is referred to as Maritime Air Superiority (MAS), which is taught in the Advanced Fighter Air Superiority (ADFAS) curriculum in the squadron. It is during ADFAS that students learn to recognize or identify Soviet and non-Soviet aircraft silhouettes so that they can employ the F-14 properly.

Subject Matter

The subject matter consisted of line drawings of front, side, and top silhouettes of Soviet and non-Soviet aircraft. A paper-based study guide was designed and developed for the subjects to help them learn to recognize silhouettes of four Soviet naval air bombers and ten of their front-line fighters. Silhouettes of non-Soviet aircraft were also presented since these could be mistaken for Soviet threats or vice versa.

The silhouettes of Soviet and non-Soviet aircraft appeared on 28 pages of the study guide. These were presented so that Soviet aircraft were displayed on a left page, and corresponding non-Soviet aircraft on the immediately following right page. A specific Soviet silhouette appeared on the left page either in the top, middle, or bottom position. The non-Soviet silhouette appeared on the right page in the corresponding top, middle, or bottom position. All top views of Soviet and non-Soviet aircraft were presented first. These were followed by all side and front views, respectively.

Each Soviet top, side, and front view had its own corresponding non-Soviet aircraft.

Subjects were asked to study each Soviet silhouette and its corresponding non-Soviet silhouette in sequence and note the distinctive features of each. The correct identification of each Soviet and non-Soviet silhouette according to NATO name and alpha-numeric designator appeared directly below it. Subjects were told that in the near future, their recognition of these Soviet and non-Soviet aircraft would be assessed via computer and traditional testing.

In addition to using the paper-based study guide, subjects were required to learn the silhouettes via the computer system described below which was configured in a training mode for this purpose. In this mode, when a student pressed the <TAB> key, a silhouette would reappear together with its correct identification so that these could be associated.

Computer-Based Assessment

Graphic models were produced to assess how well the subjects recognized or identified the above silhouettes. A computer game based upon a sequential recognition paradigm was designed and developed. It randomly selects and presents on a computer display at an arbitrary exposure setting, the front, side, or top views of four Russian bombers and 10 of their advanced fighters. For this research, the exposure of a silhouette on the computer screen was approximately 500 milliseconds. Also, the game management system can choose and flash corresponding silhouettes of NATO aircraft which act as distractors because of their high degree of similarity to the Soviet silhouettes.

This particular game, which is called FLASH IVAN (the F-14 community refers to the Russians generically as "Ivan"), assesses student performance by measuring: their "hit rate" or number of correct recognitions out of a total of 42 silhouettes half of which are Soviet and the other half non-Soviet, the time it takes a student or latency to make a recognition judgment for each target or distractor aircraft, and the degree of confidence the student has in each of his recognition decisions. At the end of the game, feedback is given to the student in terms of his hit rate (computer-based test total percentage correct responses, CTP), average response latency (computer-based test total average response latency, CTL), average degree of confidence in his recognition judgments (computer-based test total average degree of confidence, CTC), and how his performance compares to other students who have played the game.

A file is maintained and available to the instructors which provides, in addition to these parameters for each student, recognition performance across aircraft for all students who played the game. This provides diagnostic assessments to instructors who can use this summative feedback to focus student attention on learning the salient distinctive features of certain aircraft in order to improve their recognition performance.

The game management system is programmed in a modular manner: instructing the student on how to play the game, retrieving and displaying individual images, keeping track of how well students perform, providing them feedback, and linking these components in order to execute the game. This modularity in programming, together with the game management system's independence of any specific graphic

database (e.g., ship silhouettes, human anatomy, electronic circuits, topography), contributes to its wide applicability. The game, then, provides a set of software tools which can be used by others who need to assess recognition performance. This computer-based system for assessing recognition performance (RECOG) has been completely documented by Little, Maffly, Miller, Setter, & Federico (1985).

Paper-Based Assessment

Two alternative forms of a paper-based test were designed and developed to assess the subjects' recognition of the silhouettes mentioned above. The alternative test forms mimicked as much as possible the format used by FLASH IVAN. Both forms of the test were presented as booklets each containing 42 items representing the front, top, or side silhouettes of aircraft. The subjects' task was to identify as quickly as possible the aircraft that was represented by each item's silhouette. They were asked to write in the space provided what they recognized the aircraft to be (i.e., its NATO name or corresponding alphanumeric designation; e.g., FOXHOUND or MIG-31). Misspellings counted as wrong responses. Subjects were instructed not to turn back to previous pages in the test booklet to complete items they had left blank. The students were asked to go through the test items as quickly to approximate as much as possible the duration of silhouette exposure employed by FLASH IVAN. Subjects were monitored to assure they complied with this procedure.

After they wrote down what they thought an aircraft was, they were required to indicate on a scale which appeared below each silhouette the degree of confidence or sureness in their recognition decision concerning the specific item. Like the confidence scale used for FLASH IVAN, this one went from LEAST CONFIDENT or 0% CONFIDENCE in their recognition decision on the left, to MOST CONFIDENT or 100% CONFIDENCE on the right, in ten percentage point intervals. Subjects were instructed to use this confidence scale by placing a check mark directly over the percentage of confidence which best reflected or approximated the amount of sureness they had in their judgment. To learn how to respond properly to the silhouette test items, the subjects were asked to look at three completed examples. A subject's percentage of correct recognitions (paper-based test total percentage correct responses, PTP) and average degree of confidence (paper-based test total average degree of confidence, PTC) for the paper-based test were measured and recorded.

Procedure

Prior to testing, subjects learned to recognize the aircraft silhouettes using two media: (a) in paper-based form structured as a study guide, and (b) in computer-based form using FLASH IVAN in the training mode. Mode of assessment, computer-based or paper-based, was manipulated as a within-subjects variable (Kirk, 1968). All subjects were administered the paper-based test before the computer-based test. The two forms of the paper-based tests were alternated in their administration to subjects (i.e., the first subject received Form A, the second subject received Form B, the third subject received Form A, etc.). After subjects received the paper-based test, they were immediately administered the computer-based test. It was assumed that a subject's

state of recognition knowledge was the same during the administration of both tests. Subjects took approximately 10-15 minutes to complete the paper-based test, and 15-20 minutes to complete the computer-based test. This difference in completion time was primarily due to lack of typing proficiency among some of the subjects.

Reliabilities for both modes of testing were estimated by deriving internal consistency indices using an odd-even item split. These reliability estimates were adjusted by employing the Spearman-Brown Prophecy Formula (Thorndike, 1982). Reliability estimates were calculated for test score, average degree of confidence, and average response latency for the computer-based test; reliability estimates were calculated for test score and average degree of confidence only for the paper-based test. None was computed for average response latency since this was not measured for the paper-based test. Equivalences between these two modes of assessment were estimated by Pearson-product-moment correlations for total test score and average degree of confidence.

In order to derive discriminative validity estimates, research subjects were placed into two groups according to whether or not their performance through the squadron's curriculum was above or below the mean average grade for this sample. A stepwise multiple discriminant analysis, using Wilks' criterion for including and rejecting variables, and their associated statistics were computed to ascertain how well computer-based and paper-based measures distinguished among the defined groups expected to differ in their recognition of aircraft silhouettes.

Predictive validity indices were obtained by computing a canonical analysis between computer-based and paper-based recognition measures and subjects' test scores for each phase of the curriculum: (a) Familiarization Phase (FAM)--All aspects of the F-14's systems, capabilities, limitations, and emergency procedures as well as formation, instrument, night, and acrobatics flying; (b) Basic Weapons Employment (BWP)--The basics of the F-14's radar and weapon systems and rudimentary intercept procedures; (c) Guns (Gun)--The F-14's 20mm gun is taught and the trainees actually fire it at a banner towed by another aircraft and at a simulated ground target; (d) Advanced Fighter Air Superiority (ADF)--Advanced outer air battle tactics dealing with electronic counter measures emphasizing Soviet aircraft, weapons, and tactics as well as U. S. battle group tactics; and (e) Tactics (TAC)--Tactically fighting the F-14 in several likely combat scenarios against other hostile aircraft.

RESULTS

Reliability and Equivalence Estimates

Tables of reliability and equivalence estimates are presented in the appendix. Split-half reliability and equivalence estimates of computer-based and paper-based measures of recognition performance are presented in Table A-1. It can be seen that the adjusted reliability estimates are relatively high ranging from .89 to .97. The difference in reliabilities for computer-based and paper-based measures for average degree of confidence was found to be statistically significant ($p < .02$) using a test described by Edwards (1964). However, the difference in reliabilities for computer-

based and paper-based measures of the recognition test score was found to be not significant. These results revealed that (a) the computer-based and paper-based measures of test score were not significantly different in reliability or internal consistency, and (b) the paper-based measure of average degree of confidence was more reliable or internally consistent than the computer-based measure.

Equivalence estimates between corresponding computer-based and paper-based measures of recognition test score and average degree of confidence were .67 and .81, respectively. These suggested that the computer-based and paper-based measures had anywhere from approximately 45% to 66% variance in common implying that these different modes of assessment were only partially equivalent. The equivalences for test score and average degree of confidence measures were significantly ($p < .001$) different. This result suggested that computer-based and paper-based measures of average degree of confidence were more equivalent than these measures of recognition test score.

Discriminative Validity

The multiple discriminant analysis (Cooley & Lohnes, 1962; Tatsuoka, 1971; Van de Geer, 1971), which was computed to determine how well computer-based and paper-based measures of recognition performance differentiated groups defined by above or below mean average curriculum grade, yielded one significant discriminant function as expected. The statistics associated with the significant function, standardized discriminant-function coefficients, pooled within-groups correlations between the function and computer-based and paper-based measures, and group centroids for above or below mean average curriculum grade are presented in Table A-2. It can be seen that the single significant discriminant function accounted for 100% of the variance between the two groups. The discriminant-function coefficients which consider the interactions among the multivariate measures revealed the relative contribution or comparative importance of the variables in defining this derived dimension to be CTC, PTC, PTP, CTP, and CTL. The within-groups correlations which are computed for each individual measure partialling out the interactive effects of all the other variables indicated that the major contributors to the significant discriminant function were CTP, CTC, and CTL, respectively, all computer-based measures. The group centroids showed that those students whose curricular performances were above the mean average grade clustered together along one end of the derived dimension; while, those students whose curricular performances were below the mean average grade clustered together along the other end of the continuum.

The means and standard deviations for groups above or below mean average curriculum grade, univariate F-ratios, and levels of significance for computer-based and paper-based measures of recognition performance are tabulated in Table A-3. Considering the measures as univariate variables (i.e., independent of their multivariate relationships with one another) these statistics revealed that one computer-based measure, CTL, and one paper-based measure, PTC, significantly differentiated the two groups. The means revealed that the group above mean average curriculum grade had shorter computer-based latencies than the group below mean average curriculum grade, and that the former group had a higher paper-based average degree of confidence than the

latter group. In general, the multivariate and subsequent univariate results established that according to two sets of criteria, the discriminant coefficients and F-ratios and corresponding means, the discriminant validities of computer-based and paper-based measures were about the same for distinguishing groups above or below mean average curriculum grade. However, according to another set of criteria, the pooled within-groups correlations between the discriminant function and the computer-based and paper-based measures, the former had superior discriminative validity than the latter.

Predictive Validity

The statistics associated with the significant canonical correlation (Cooley & Lohnes, 1962) between computer-based and paper-based measures of recognition performance and curricular criteria are presented in Table A-4. These results established that the computer-based and paper-based measures of recognition performance were significantly associated with the curricular criteria. The canonical variates revealed that the major contributors to this correlation in order of importance were PTC, CTC, CTL, BWP, and ADF. When the relative magnitudes of the canonical variates are considered, the paper-based measure, PTC, is the most salient contributor to the correlation. However, 50 percent of the paper-based measures and 66 percent of the computer-based measures were the primary contributors to the multivariate relationship between recognition performance and the basic weapons and advanced fighter air superority phases of of the curriculum. The univariate relationships among the above five major contributors to the canonical correlation as reflected by the Pearson product-moment correlations revealed that CTL and PTC were significantly associated with BWP and ADF. Nevertheless, the differences in the strength of the associations of CTL and PTC with BWP and ADF were found to be not significantly different. All of these statistics associated with the canonical correlation suggested the predictive validity of computer-based measures approximates that of paper-based measures.

DISCUSSION

This study established that the relative reliability of computer-based and paper-based measures depends upon the specific criterion assessed. That is, regarding the recognition test score itself, it was found that computer-based and paper-based measures were not significantly different in reliability or internal consistency. However, regarding the average degree of confidence in recognition judgments, it was found that the paper-based measure was more reliable or internally consistent than its computer-based counterpart. The extent of the equivalence between these two modes of measurement was contingent upon particular performance criteria. It was demonstrated that the equivalence of computer-based and paper-based measures of average degree of confidence was greater than that for recognition test score. The relative discriminative validity of computer-based and paper-based measures was dependent upon the specific statistical criteria selected. The discriminant coefficients, F-ratios, and corresponding means indicated that the validities of computer-based and paper-based measures were about the same for distinguishing groups above or below mean average curriculum grade. However, according to another set of criteria, the pooled within-groups

correlations between the discriminant function and computer-based and paper-based measures, the former had superior validity than the latter. Also, according to the statistics associated with the canonical correlation, this study demonstrated that the predictive validity of computer-based measures approximates that of paper-based measures. The results of this research supported the findings of some studies, but not others. As will be discussed, the reported literature on this subject is contradictory and inconclusive.

Federico and Liggett (1988, 1989) administered computer-based and paper-based tests of threat-parameter knowledge (Liggett & Federico, 1986) in order to determine the relative reliability and validity of these two modes of assessment. Estimates of internal consistencies, equivalences, and discriminant validities were computed. They established that computer-based and paper-based measures (i.e., test score and average degree of confidence) were not significantly different in reliability or internal consistency. This finding partially agrees with the corresponding result of this present study since computer-based and paper-based measures of test score were found to be equally reliable; however, the computer-based measure of average degree of confidence was found to be less reliable than its paper-based counterpart. A few of the Federico and Liggett findings were ambivalent since some results suggested equivalence estimates for computer-based and paper-based measures (i.e., test score and average degree of confidence) were about the same, and another suggested these estimates are different. Some of this reported result is different from that established in this present study where computer-based and paper-based measures of test score were less equivalent than these measures of average degree of confidence. Lastly, Federico and Liggett demonstrated that the discriminative validity of the computer-based measures was superior to paper-based measures. This result is in partial agreement with that found in this reported research where this was also established with respect to some statistical criteria. However, according to other criteria, the discriminative validity of computer-based and paper-based measures were about the same.

Hofer and Green (1985) were concerned that computer-based assessment would introduce irrelevant or extraneous factors that would likely degrade test performance. These computer-correlated factors may alter the nature of the task to such a degree, it would be difficult for a computer-based test and its paper-based counterpart to measure the same construct or content. This could impact upon reliability, validity, normative data, as well as other assessment attributes. Several plausible reasons, they stated, may contribute to different performances on these distinct kinds of testing: (a) state anxiety instigated when confronted by computer-based testing, (b) lack of computer familiarity on the part of the testee, and (c) changes in response format required by the two modes of assessment. These different dimensions could result in tests that are nonequivalent; however, in this reported research these diverse factors had no apparent impact.

On the other hand, there are a number of known differences between computer-based and paper-based assessment which may affect equivalence and validity: (a) Passive omitting of items is usually not permitted on computer-based tests. An individual must respond unlike most paper-based tests. (b) Computerized tests typically do not permit backtracking. The testee cannot easily review items, alter responses, or delay attempting to answer questions. (c) The capacity of the computer screen can have an

impact on what usually are long test items, e.g., paragraph comprehension. These may be shortened to accommodate the computer display, thus partially changing the nature of the task. (d) The quality of computer graphics may affect the comprehension and degree of difficulty of the item. (e) Pressing a key or using a mouse is probably easier than marking an answer sheet. This may impact upon the validity of speeded tests. (f) Since the computer typically displays items individually, traditional time limits are no longer necessary (Green, 1986).

Sampson (1983) discussed some of the potential problems associated with computer-based assessment: (a) not taking into account human factors principles to design the human-computer interface, (b) individuals may become anxious to such a degree when having to interact with a computer for assessment that the measurement obtained may be questionable, (c) unauthorized access and invasion of privacy are just some of the abuses that can result from computerized testing, (d) inaccurate test interpretations by users of the system can easily culminate in erroneously drawn conclusions, (e) differences in modes of administration may make paper-based norms inappropriate for computer-based assessment, (f) lack of reporting reliability and validity data for computerized tests, and (g) resistance toward using new computer-based systems for performance assessment. A potential limitation of computer-based assessment is depersonalization and decreased opportunity for observation. This is especially true in clinical environments (Space, 1981). Most computer-based tests do not allow individuals to omit or skip items, or to alter earlier responses. This procedure could change the test-taking strategy of some examinees. To permit it, however, would probably create confusion and hesitation during the process of retracing through items as the testee uses clues from some to minimize the degree of difficulty of others (Green, Bock, Humphreys, Linn, & Reskase, 1984).

Some of the comments made by Colvin and Clark (1984) concerning instructional media can be easily extrapolated to assessment media. (Training and testing are inextricably intertwined; it is difficult to do one well without the other.) This is especially appropriate regarding some of the attitudes and assumptions permeating the employment of, and enthusiasm for, media: (a) confronted with new media, computer-based or otherwise, students will not only work harder, but also enjoy their training and testing more; (b) matching training and testing content to mode of presentation is important, even though not all that prescriptive or empirically well established; (c) the application of computer-based systems permits self-instruction and self-assessment with their concomitant flexibility in scheduling and pacing training and testing; (d) monetary and human resources can be invested in designing and developing computer-based media for instruction and assessment that can be used repeatedly and amortized over a longer time, rather than in labor intensive classroom-based training and testing; and (e) the stability and consistency of instruction and assessment can be improved by media, computer-based or not, for distribution at different times and locations however remote.

When evaluating or comparing different media for instruction and assessment, the newer medium may simply be perceived as being more interesting, engaging, and challenging by the students. This novelty effect seems to disappear as rapidly as it appears. However, in research studies conducted over a relatively short time span, e.g., a few days or months at the most, this effect may still be lingering and affecting the

evaluation by enhancing the impact of the more novel medium (Colvin & Clark, 1984). When matching media to distinct subject matters, course contents, or core concepts, some research evidence (Jamison, Suppes, & Welles, 1974) indicates that, other than in obvious cases, just about any medium will be effective for different content.

Another salient question that should be addressed is: How to combine effectively and efficiently computer and cognitive science, artificial intelligence (AI) technology, current psychometric theory, and diagnostic testing? It has been demonstrated (Brown & Burton, 1978; Kieras, 1987; McArthur & Choppin, 1984; Wenger, 1987) that AI techniques can be developed to diagnose specific error-response patterns or bugs to advance measurement methodology.

RECOG together with the Soviet and non-Soviet aircraft silhouette database is referred to as FLASH IVAN. This system is currently being used to augment the teaching and testing of this subject matter in VF-124. RECOG was designed and developed with generalizability (i.e., independence of subject-matter domain) and transferability (i.e., capable of readily running on different computer systems) in mind as was the Computer-Based Educational Software System (CBESS) (Brandt, 1987; Brandt, Gay, Othmer & Halff, 1987). CBESS consists of a number of component quizzes such as JEOPARDY, TWENTY QUESTIONS, and CONSTRAINT. Since the time that RECOG and FLASH IVAN were developed and evaluated, two other tests, FLASH and PICTURE, were added to CBESS. In terms of their function, these additional quizzes are similar to RECOG and FLASH IVAN. However, these more recently produced tests are written in the "C" programming language for the Navy standard microcomputer, Zenith Z-248. Consequently, since TERAK computers are no longer being produced, the company went out of business, FLASH and PICTURE can be perceived as replacements for RECOG and FLASH IVAN.

RECOMMENDATIONS

Based upon the findings of this study, the following actions are recommended:

(a) Commander, Naval Air Force, U.S. Pacific Fleet, use FLASH and PICTURE to supplement the training and testing of fighter and other crew members to recognize Soviet and Non-Soviet silhouettes.

(b) Chief of Naval Operations, Total Force Training and Education, fund the evaluation and seek implementation of FLASH and PICTURE in other content areas or subject-matter domains (e.g., ship silhouettes, electronic schemata, human anatomy) to ascertain the universality of the validity and reliability results established in this reported research.

REFERENCES

Brandt, R. C. (1987). *Computer-based memorization system (CBMS): Student manual.* Salt Lake City: Department of Computer Science, University of Utah.

Brandt, R. C., Gay, L. S., Othmer, B., & Halff, H. M. (1987). *Computer-based memorization system (CBMS): Author and instructor manual.* Salt Lake City: Department of Computer Science, University of Utah.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in mathematical skills. *Cognitive Science, 2,* 155-192.

Colvin, C., & Clark, R. E. (1984). Instructional media vs. instructional methods. *Performance and Instruction Journal,* July, 1-3.

Cooley, W. W., & Lohnes, P. R. (1962). *Multivariate procedures for the behavioral sciences.* New York: John Wiley & Sons.

Cory, C. H. (1977). Relative utility of computerized versus paper-and-pencil tests for predicting job performance. *Applied Psychological Measurement, 1,* 551-564.

Divgi, D. R. (1988, October). *Two consequences of improving a test battery* (CRM 88-171). Alexandria VA: Center for Naval Analyses.

Edwards, A. L. (1964). *Experimental design in psychological research.* New York: Holt, Rinehart, and Winston.

Elwood, D. L., & Griffin, R. H. (1972). Individual intelligence testing without the examiner: Reliability of an automated method. *Journal of Consulting and Clinical Psychology, 38,* 9-14.

English, R. A., Reckase, M. D., & Patience, W. M. (1977). Applications of tailored testing to achievement measurement. *Behavior Research Methods & Instrumentation, 9,* 158-161.

Evan, W. M., & Miller, J. R. (1969). Differential effects of response bias of computer versus conventional administration of a social science questionnaire. *Behavioral Science, 14,* 216-227.

Federico, P-A. (1980). Adaptive instruction: Trends and issues. In R. E. Snow, P-A. Federico, & W. E. Montague (Eds.), *Aptitude, learning, and instruction, volume 1: Cognitive process analyses of aptitude.* Hillsdale NJ: Erlbaum.

Federico, P-A., & Liggett, N. L. (1988, April). *Comparing computer-based and paper-based assessment strategies for semantic knowledge*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Federico, P-A., & Liggett, N. L. (1989). *Computer-based and paper-based measurement of semantic knowledge* (NPRDC TR 89-4). San Diego: Navy Personnel Research and Development Center.

Green, B. F. (1983a). Adaptive testing by computer. *Measurement, Technology, and Individuality in Education. 17*, 5-12.

Green, B. F. (1983b). The promise of tailored tests. *Principles of modern psychological measurement: A festschrift in honor of Frederic Lord*. Hillsdale NJ: Erlbaum.

Green, B. F. (1986). *Construct validity of computer-based tests*. Paper presented at the test validity conference educational testing service, Princeton, N. J.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Hansen, D. H., & O'Neil, H. F. (1970). Empirical investigations versus anecdotal observations concerning anxiety and computer-assisted instruction. *Journal of School Psychology, 8*, 315-316.

Hedl, J. J., O'Neil, H. F., & Hansen, D. H. (1973). Affective reactions toward computer-based intelligence testing. *Journal of Consulting and Clinical Psychology, 40*, 217-222.

Hitti, F. J., Riffer, R. L., & Stuckless, E. R. (July, 1971). *Computer-managed testing: A feasibility study with deaf students*. National Technical Institute for the Deaf.

Hofer, P. J., & Green, B. F. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting and Clinical Psychology, 53*, 826-838.

Hoffman, K. I., & Lundberg, G. D. (1976). A comparison of computer-monitored group tests with paper-and-pencil tests. *Educational and Psychological Measurement, 36*, 791-809.

Jamison, D., Suppes, P., & Welles, S. (1974). The effectiveness of alternative media:

A survey. *Annual Review of Educational Research, 44,* 1-68.

Johnson, J. H., & Johnson, K. N. (1981). Psychological considerations related to the development of computerized testing stations. *Behavior Research Methods & Instrumentation, 13,* 421-424.

Johnson, D. F., & Mihal, W. L. (1973). Performance of black and whites in computerized versus manual testing environments. *American Psychologist, 28,* 694-699.

Johnson, D. F., & White, C. B. (1980). Effects of training on computerized test performance in the elderly. *Journal of Applied Psychology, 65,* 357-358.

Kantor, J. (1988). *The effects of anonymity, item sensitivity, trust, and method of adminstration on response bias on the job description index.* Unpublished doctoral dissertation, California Schoool of Professional Psychology, San Diego.

Katz, L., & Dalby, J. T. (1981). Computer-assisted and traditional psychological assessment of elementary-school-age children. *Contemporary Educational Psychology, 6,* 314-322.

Kieras, D. E. (1987). *The role of cognitive simulation models in the development of advanced training and testing systems* (TR-87/ONR-23). Ann Arbor: University of Michigan.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences.* Belmont CA: Brooks/Cole.

Koson, D., Kitchen, C., Kochen, M., & Stodolosky, D. (1970). Psychological testing by computer: Effect on response bias. *Educational and Psychological Measurement, 30,* 808-810.

Lee, J. A., Moreno, K. E., & Sympson, J. B. (April, 1984). *The effects of mode of test administration on test performance.* Paper presented at the annual meeting of the Eastern Psychological Association, Baltimore.

Liggett, N. L., & Federico, P-A. (1986). *Computer-based system for assessing semantic knowledge: Enhancements* (NPRDC TN 87-4). San Diego: Navy Personnel Research and Development Center.

Little, G. A., Maffly, D. H., Miller, C. L., Setter, D. A., & Federico, P-A. (1985). *A computer-based gaming system for assessing recognition performance (recog)* (TL 85-3). San Diego, California: Training Laboratory, Navy Personnel Research and Development Center.

Lucas, R. W., Mullin, P. J., Luna, C. D., & McInroy, D. C. (1977). Psychiatrists and a computer as interrogators of patients with alcohol related illnesses: A comparison. *British Journal of Psychiatry, 131,* 160-167.

Lukin, M. E., Dowd, E. T., Plake, B. S., & Kraft, R. G. (1985). Comparing computerized versus traditional psychological assessment. *Computers in Human Behavior, 1,* 49-58.

Lushene, R. E., O'Neil, H. F., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment, 34,* 353-361.

McArthur, D. L., & Choppin, B. H. (1984). Computerized diagnostic testing. *Journal of Educational Measurement, 21,* 391-397.

McBride, J. R. (1980). Adaptive verbal ability testing in a military setting. In D. J. Weiss (Ed.), *Proceedings of the 1979 computerized adaptive testing conference.* Minneapolis: University of Minnesota, Department of Psychology.

Millman, J. (1984). Using microcomputers to administer tests: An alternate point of view. *Educational Measurement: Issues and Practices,* Summer, 20-21.

Rock, D. L., & Nolen, P. A. (1982). Comparison of the standard and computerized versions of the raven coloured progressive matrices test. *Perceptual and Motor Skills, 54,* 40-42.

Sachar, J. D., & Fletcher, J. D. (1978). Administering paper-and-pencil tests by computer, or the medium is not always the message. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology.

Sampson, J. R. (1983). Computer-assisted testing and assessment: Current status and implications for the future. *Measurement and Evaluation in Guidance, 15,* 293-299.

Serwer, B. L., & Stolurow, L. M. (1970). Computer-assisted learning in language arts. *Elementary English, 47,* 641-650.

Skinner, H. A., & Allen, B. A. (1983). Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. *Journal of Consulting and Clinical Psychology, 51,* 267-275.

Space, L. G. (1981). The computer as psychometrician. *Behavior Research Methods & Instrumentation, 13,* 595-606.

Sympson, J. B., Weiss, D. J., & Ree, M. (1982). *Predictive validity of conventional and adaptive tests in an air force training environment* (AFHRL-TR-81-40). Brooks AFB: Air Force Human Resources Laboratory.

Tatsuoka, M. M. (1971). *Multivariate analysis*. New York: John Wiley & Sons.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

Van de Geer, J. P. (1971). *Introduction to multivariate analysis for the social sciences*. San Francisco: W. H. Freeman.

Ward, W. C. (1984). Using microcomputers to administer tests. *Educational Measurement: Issues and Practices*, Summer, 16-20.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge*. Los Altos, CA: Kaufman.

Wildgrube, W. (July, 1982). *Computerized testing in the german federal armed forces--empirical approaches*. Paper presented at the 1982 Computerized Adaptive Testing Conference, Spring Hill MN.

APPENDIX

TABLES OF RELIABILITY AND VALIDITY ESTIMATES

Page

A-1. Split-Half Reliability and Equivalence Estimates of Computer-Based and Paper-Based Measures of Recognition Performance A-1

A-2. Statistics Associated with Significant Discriminant Function, Standardized Discriminant-Function Coefficients, Pooled Within-Groups Correlations Between the Discriminant Function and Computer-Based and Paper-Based Measures, and Group Centroids for Above or Below Mean Average Curriculum Grade A-2

A-3. Means and Standard Deviations for Groups Above or Below Mean Average Grade, Univariate F-Ratios, and Levels of Significance for Computer-Based and Paper-Based Measures A-3

A-4. Statistics Associated with Significant Canonical Correlation Between Computer-Based and Paper-Based Measures of Recognition Performance A-4

Table A-1

Split-Half Reliability and Equivalence Estimates of Computer-Based and Paper-Based Measures of Recognition Performance

Measure	Reliability		Equivalence
	Computer-Based	Paper-Based	
Score	.90	.89	.67
Confidence	.95	.97	.81
Latency	.93	--	--

Note. Split-half reliability estimates were adjusted by employing the Spearman-Brown Prophecy Formula.

Table A-2

Statistics Associated with Significant Discriminant Function, Standardized Discriminant-Function Coefficients, Pooled Within-Groups Correlations Between the Discriminant Function and Computer-Based and Paper-Based Measures, and Group Centroids for Above or Below Mean Average Curriculum Grade

Discriminant Function

Eigen-value	Percent Variance	Canonical Correlation	Wilks Lambda	Chi Squared	d.f.	p
.14	100.00	.35	.88	9.98	5	.076

Measure	Discriminant Coefficient	Within-Group Correlation
CTP	.60	-.60
CTC	.97	.55
CTL	.52	-.48
PTP	-.80	-.25
PTC	-.94	.03

Group	Centroid
Above Mean Average Grade	-.32
Below Mean Average Grade	.42

Table A-3

Means and Standard Deviations for Groups Above or Below Mean Average Grade, Univariate F-Ratios, and Levels of Significance for Computer-Based and Paper-Based Measures

Measure		Group: Above Mean Average Grade (n=47)	Group: Below Mean Average Grade (n=36)	F	p
CTP	$\bar{X}$	77.19	77.64	.01	.92
	s	18.48	23.33		
CTC	$\bar{X}$	90.99	88.54	.73	.39
	s	12.74	13.06		
CTL	$\bar{X}$	1522.06	2115.61	3.31	.07
	s	1554.12	1359.19		
PTP	$\bar{X}$	86.40	80.42	2.56	.11
	s	16.65	17.19		
PTC	$\bar{X}$	94.09	89.61	3.92	.05
	s	9.47	11.09		

Table A-4

Statistics Associated with the Significant Canonical Correlation Between Computer-Based and Paper-Based Measures of Recognition Performance and Curricular Criteria.

Canonical Correlation	Eigen-Value	Wilks Lambda	Chi Squared	d.f.	p
.51	.26	.55	45.13	25	.008

	Computer-Based and Paper-and-Pencil Measures	Canonical Variate	Curricular Criteria	Canonical Variate
	CTP	.28	FAM	-.11
	CTC	-1.06	BWP	.65
	CTL	-.73	GUN	.02
	PTP	-.21	TAC	-.02
	PTC	1.17	ADF	.54

Pearson Product-Moment Correlations

	BWP	ADF
CTC	.08	.15
CTL	-.30[a]	-.35[a]
PTC	.29[a]	.27[a]

Note: a. $r(81) > .256$; $p < .01$.

DISTRIBUTION LIST

Distribution:
Assistant for Manpower Personnel and Training Research and Development (OP-01B2)
Chief of Naval Operations (OP-11B1)
Head, Training and Education Assessment (OP-11H)
Technology Area Manager, Office of Naval Technology (Code 222)
Commander, Naval Air Force, U.S. Pacific Fleet (Code 313)
Defense Technical Information Center (DTIC) (2)

Copy to:

Cognitive and Decision Science (OCNR-1142CS)
Naval Military Personnel Command, Library (Code NMPC-013D)
Technical Director, U.S. ARI, Behavioral and Social Sciences, Alexandria, VA (PERI-ZT)
TSRL/Technical Library (FL 2870)
Superintendent, Naval Postgraduate School
Director of Research, U.S. Naval Academy
Institute for Defense Analyses, Science and Technology Division
Center for Naval Analyses, Acquisitions Unit